IET Signal Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Bandwidth selection for kernel density estimation using Fourier domain constraints

ISSN 1751-9675
Received on 17th May 2015
Revised on 17th October 2015
Accepted on 1st December 2015
doi: 10.1049/iet-spr.2015.0076
www.ietdl.org

Alexander Suhre ✉, *Orhan Arikan*, *Ahmed Enis Cetin*

Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
✉ *E-mail: alexander.suhre@gmail.com*

Abstract: Kernel density estimation (KDE) is widely-used for non-parametric estimation of an underlying density from data. The performance of KDE is mainly dependent on the bandwidth parameter of the kernel. This study presents an alternative method of estimating the bandwidth by incorporating sparsity priors in the Fourier transform domain. By using cross-validation (CV) together with an l_1 constraint, the proposed method significantly reduces the under-smoothing effect of traditional CV methods. A solution for all free parameters in the minimisation is proposed, such that the algorithm does not need any additional parameter tuning. Simulation results indicate that the new approach is able to outperform classical and more recent approaches over a set of distributions of interest.

1 Introduction

Estimating an underlying distribution from data is a widely studied problem [1]. Probability density function (PDF) estimation approaches can broadly be divided into two classes: parametric estimation and non-parametric estimation. An important branch of non-parametric estimation is the kernel-based approach, which is frequently referenced as kernel density estimation (KDE) [2]. In this approach, an estimate for the underlying density $g_X(x)$ is given by

$$\hat{g}_X(x;\sigma) = \frac{1}{N}\sum_{i=0}^{N-1} k_\sigma(x - v_i), \tag{1}$$

where N is the number of data points, v_i with $i = 0, 1, 2, \ldots, N$ denotes the observed data and σ is called the bandwidth of the kernel k_σ, which corresponds to the standard deviation for Gaussian kernels.

The performance of KDE depends largely on the bandwidth of the kernel, which, if not chosen appropriately, can result in an over-smoothed estimate, i.e. containing little detail and therefore having small support in the Fourier domain, or an under-smoothed estimate, i.e. containing a lot of detail and therefore having a large support in the Fourier domain. The chosen bandwidth should decrease the mean integrated square error (MISE) [3]. In [4] several techniques for bandwidth estimation have been evaluated and it has been concluded that the most efficient method is the 'plug-into-equation' approach of Sheather and Jones [5], Raykar *et al.* [6], where the bandwidth σ is determined by minimising the MISE [3]. For mathematical tractability, minimisation of the approximate MISE (AMISE) is carried out in [3], where the optimal σ is chosen based on the second derivative of $g_X(x)$. More recent methods include Botev's method [7], where the used plug-in method is free from the normal reference rule.

There are also other ways of estimating the bandwidth. Silverman [2], for example discusses 'leave-one-out' cross-validation (CV) as another tool. In [4], this method was evaluated as somehow inferior to the 'plug-into-equation' approaches, therefore the general opinion in the statistics community is that CV-type approaches tend to produce under-smoothed estimates of a distribution. However, Loader [8] is at odds with this broad categorisation, stating that 'the comparisons between classical and plug-in approaches presented in the literature have several weaknesses. First, plug-in approaches, through the specification of tuning parameters for pilot estimates, effectively make substantial prior assumptions about the required bandwidth and will fail if this information is wrong. Second, the plug-in approaches obtain much of their information from the data through the use of higher order pilot estimates; if classical approaches are also allowed to consider higher order methods, better estimates result. Third, plug-in methods are not rescued by asymptotic analysis showing better rates of convergence; assumptions about the underlying function make the resulting estimate asymptotically inefficient, regardless of how good the bandwidth selector is.

Our proposed method was developed with Loader's arguments in mind. We want to use CV, since it will not make faulty prior assumptions. However, we need to reduce its tendency to under-smooth. In this paper, we propose a method for estimating the bandwidth of the kernel by minimising a new cost function consisting of the CV term and an l_1 constraint implemented in the Fourier domain. The computational cost of implementing the estimator in (1), is $\mathcal{O}(N^2)$ multiplications, which for large datasets may be prohibitive. A significant number of studies have been devoted to KDE, especially with the goal of reducing its computational burden [9, 6, 10]. In [10], (1) is implemented in the Fourier domain via multiplication. The order of solving (1) in the Fourier domain is then $\mathcal{O}(N\log(N))$ by using the fast Fourier transform (FFT) algorithm. The proposed method takes advantage of the sparsity of the data in the Fourier domain.

In the following, we review the CV method as proposed by Silverman [2] and propose a new cost function for CV that includes an l_1 term. We then present simulation results.

2 CV-based cost function for bandwidth estimation

In [2], the least-squares CV was introduced as a way to find an estimate $\hat{g}_x(x)$ that minimises the integrated square error (ISE)

$$\begin{aligned}\mathrm{ISE}(\hat{g}_X(x)) &= \int \left(\hat{g}_X(x) - g_X(x)\right)^2 \mathrm{d}x \\ &= \int \hat{g}_X^2(x)\mathrm{d}x - 2\int \hat{g}_X(x) g_X(x)\mathrm{d}x + \int g_X^2(x)\mathrm{d}x. \end{aligned}\tag{2}$$

Since the last term in (2) does not depend on the data, it is sufficient to find an estimate that minimises the first two terms of the ISE,

denoted by

$$R\big(\hat{g}_X(x)\big) = \int \hat{g}_X^2(x)\mathrm{d}x - 2\int \hat{g}_X(x)\cdot g_X(x)\mathrm{d}x. \tag{3}$$

According to [2], (3) can be estimated by

$$M_0(\sigma) = \int \hat{g}_X^2(x)\mathrm{d}x - \frac{2}{N}\sum_{l=0,l\neq i}^{N} \hat{g}_{X-i}(v_l), \tag{4}$$

where $\hat{g}_{X-i}(v_l)$ denotes the discrete density estimate constructed from all the datapoints except the ith observation v_i.

In the CV approach, the bandwidth σ is estimated by minimising $M_0(\sigma)$. In the following, the problem is stated in the Fourier domain: let $g_x(x)$ denotes the original distribution from which N samples are drawn independently. Equation (1) can be written as a convolution as follows

$$\hat{g}_x(x;\,\sigma) = k_\sigma(x) \times \frac{1}{N}\sum_{i=0}^{N-1}\delta(x - v_i). \tag{5}$$

It is straight forward to see that the Fourier transform of (5) is

$$\hat{G}_x(\omega;\,\sigma) = K_\sigma(\omega)\cdot\frac{1}{N}\sum_{i=0}^{N-1}\mathrm{e}^{-\mathrm{j}\omega v_i} = K_\sigma(\omega)\cdot\hat{H}(\omega), \tag{6}$$

where K_σ and $\hat{H}(\omega)$ are the Fourier transforms of the kernel k_σ and the data, respectively. Implementation of (6) is carried out using the discrete Fourier transform (DFT). A discrete estimate $\hat{H}[k]$ of $\hat{H}(\omega)$ can easily be obtained by using uniform binning of the data in the interval $[-L,\,L]$ into N intervals and computing the FFT of the binned data, which is the histogram $\hat{h}[i]$. If the kernel $k_\sigma(x)$ is chosen to be a standard normal Gaussian function with zero mean and variance σ, i.e. $k_\sigma(x) = 1/\sqrt{2\pi}\sigma\mathrm{e}^{-x^2/2\sigma^2}$, its Fourier transform $K_\sigma(\omega)$ is again a Gaussian function and can be written according to [11] as

$$K_\sigma(\omega) = \mathrm{e}^{-2\pi^2\sigma^2\omega^2}, \tag{7}$$

which is also discretised in the DFT-based implementation. In the proposed approaches, σ will be estimated as the minimiser of a new cost function that includes the l_1 norm of the Fourier transform of $\hat{g}_X(x;\,\sigma)$.

3 CV using l_1 norm in Fourier domain

As mentioned above, the least-squares CV tends to under-smooth its estimate of a distribution. Under-smoothing means that the Fourier transform of the estimate has large support in Fourier domain. Therefore, one would prefer estimates that are somewhat sparse in the Fourier domain. A large body of the literature on sparsity exists, e.g. in [12]. Sparsity in the Fourier domain can be achieved by minimising the l_0 norm of the DFT coefficients. While minimising the l_0 norm of a signal is non-deterministic polynomial (NP)-hard (NP time), we can approximate it by the l_1 norm, where the minimisation is far easier to carry out. Taking this into account, we propose the new cost function as follows

$$\min_{\sigma}\ \lambda\cdot M_0(\sigma) + (1-\lambda)\cdot|\mathcal{F}\{\hat{g}_X(x)\}|_1, \tag{8}$$

where the mixture parameter λ takes values between 0 and 1 and the second term denotes the l_1 norm of the DFT of $\hat{g}_X(x)$ as follows

$$|\mathcal{F}\{\hat{g}_X(x)\}|_1 = \sum_{k=0}^{K/2-1}|G(k)|, \tag{9}$$

where K is the DFT size and $G(k)$ denotes the DFT coefficients.

The parameter λ in (8) is the linear combination parameter. Since the first term is in the sample domain and the second term is in the Fourier domain, they have different proportions. We want them to contribute to the overall cost function (8) in an approximately equal manner. Therefore, the remaining problem is now to find a suitable λ. One can carry out a greedy search to find suitable parameter values, but this would make the method computationally inefficient. We are interested in finding an estimate for λ from the data. Let us rewrite the expression from (8) that is to be minimised, as the following convex cost function

$$\begin{aligned} C(\lambda;\,\sigma) &= \lambda\cdot M_0(\sigma) + (1-\lambda)\cdot|\mathcal{F}\{\hat{g}_X(x)\}|_1 \\ &= \lambda\cdot J_1(\sigma) + (1-\lambda)\cdot J_2(\sigma) \end{aligned} \tag{10}$$

We want to find the λ that minimises C from (10). However, the choice of λ depends on the values of J_1 and J_2 and therefore on the choice of σ. A simple example will illustrate this point: consider a value σ_a, where $J_1(\sigma_a) = 1$ and $J_2(\sigma_a) = 0$. The optimal choice for λ in this case is 1. Conversely, in another case with a value of σ_b, where $J_1(\sigma_b) = 0$ and $J_2(\sigma_b) = 1$, the optimal choice for λ is 0. Our point here is that even in less extreme cases, one has to consider the dependency of λ on σ.

Our proposed method for finding λ is explained in Fig. 1. This figure plots $J_2(\sigma)$ versus $J_1(\sigma)$. Each tangent to the curve represents an equipotential cost line, i.e. each point on a tangent to the curve has the same C according to (10). Therefore, the cost C of the tangential point of the curve is equivalent to the gradient. We want to choose the point with the smallest cost, so we need to find the point with the smallest gradient magnitude. To find our desired λ, we need to set $\delta C(\lambda;\,\sigma)/\sigma = 0$, yielding

$$\frac{\delta J_2(\sigma)}{\delta J_1(\sigma)} = \frac{\delta J_2(\sigma)/\delta\sigma}{\delta J_1(\sigma)/\delta\sigma} = -\frac{\lambda}{1-\lambda}. \tag{11}$$

The desired parameter λ can now be found easily. The terms $\delta J_2(\sigma)/\delta\sigma$ and $\delta J_1(\sigma)/\delta\sigma$ are easy to compute in a discrete implementation by using the differences between consecutive J_1 or J_2 values, respectively.

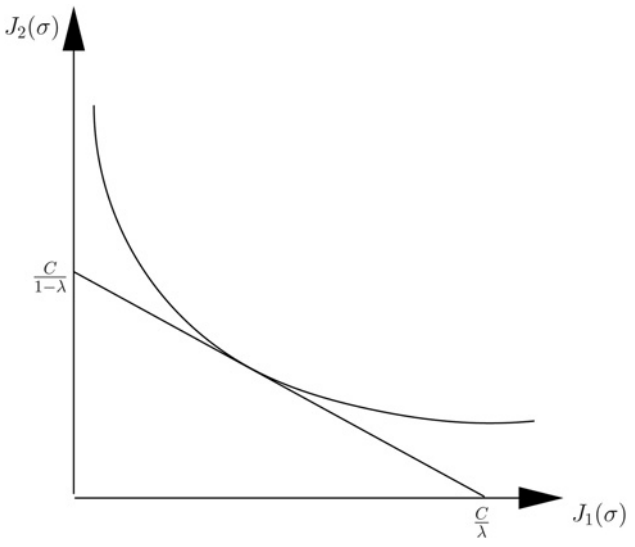

Fig. 1 *Example plot of our proposed cost minimisation method for choosing λ. We want to find the λ that corresponds to the tangent with the smallest gradient*

Table 1 KL divergence gain in dB over Sheather's method for traditional CV, Botev's and proposed methods, N was chosen as 128 and 256

Number	$N=128$			$N=256$		
	CV	Botev	Ours	CV	Botev	Ours
1	−25.48	−0.39	−3.04	−29.48	−0.32	−2.50
2	−23.40	−0.43	−2.86	−27.46	−0.13	−2.85
3	−6.93	2.60	3.34	−11.52	2.48	2.57
4	−13.37	−0.87	−2.16	−17.14	−0.93	−1.96
5	−5.69	−0.08	−0.36	−9.22	−0.19	−0.82
6	−24.22	−2.21	−0.29	−27.77	−1.59	−0.54
7	−25.26	−5.85	−0.80	−28.08	−0.43	−0.85
8	−21.98	−1.71	−0.50	−25.89	−1.09	−0.90
9	−23.76	−2.22	−0.11	−26.65	−1.41	0.00
10	−10.50	0.41	1.43	−11.15	4.50	4.72
11	−24.78	−2.22	−0.85	−26.65	−1.34	−0.69
12	−12.33	−0.37	2.62	−14.04	1.36	2.66
13	−22.98	−1.65	−0.96	−24.20	−0.92	−0.88
14	7.40	−3.78	6.95	4.65	3.90	8.08
15	−1.15	−4.07	3.60	−2.60	1.61	6.28
mean	−15.63	−1.52	0.42	−18.48	0.37	0.82

Table 2 KL divergence in dB over Sheather's method for traditional CV, Botev's and proposed methods, N was chosen as 512 and 1024

Number	$N=512$			$N=1024$		
	CV	Botev	Ours	CV	Botev	Ours
1	−31.21	−0.25	−2.75	−30.05	−0.15	−2.51
2	−29.07	−0.09	−2.60	−27.78	−0.03	−2.43
3	−14.22	1.96	1.86	−13.28	1.19	1.07
4	−18.22	−0.90	−1.52	−16.52	−0.84	−1.19
5	−11.67	−0.19	−0.61	−11.96	−0.15	−0.25
6	−30.89	−0.93	−0.91	−28.11	−0.46	−1.14
7	−28.58	−0.25	−0.85	−26.61	−0.17	−0.87
8	−28.17	−0.48	−0.97	−25.94	−0.22	−0.98
9	−28.98	−0.76	−0.30	−25.87	−0.32	−0.39
10	−10.49	6.89	6.52	−14.85	1.01	0.78
11	−28.34	−0.67	−0.73	−14.28	−0.34	−0.92
12	−14.80	1.48	2.22	−10.39	1.62	2.52
13	−25.32	−0.42	−0.83	−20.02	−0.08	−0.06
14	6.97	5.80	9.30	8.10	7.67	9.89
15	0.96	4.71	9.00	2.31	11.25	10.05
mean	−19.47	1.06	1.12	−17.68	1.33	0.91

4 Simulation results

The performance of the proposed methods is evaluated by using 15 test distributions from [3]. These distributions are mixtures of Gaussians of different flavours. Some of the original PDFs are smooth and unimodal, some of them are multimodal and have sharp peaks. N random variates were independently drawn from each of the distributions. The performances of these methods were measured against the original test distributions using the Kullback–Leibler (KL) divergence as error criteria.

In the experiments, the influence of different choices of N on the performance of the proposed method was investigated. N was varied between 2^7, 2^8, 2^9, and 2^{10}. For each N, the experiments were again carried out 500 times for each distribution and results were averaged. Results can be seen in Table 1 for $N=2^7$, 2^8 and Table 2 for $N=2^9$, 2^{10}. In these tables, the gain in dB over the Sheather–Jones method is given for different distributions, different N values and three methods that were compared: the traditional CV method, Botev's method and the proposed method. Let M_s denotes the value of Kullback–Leibler (KL) divergence for the Sheather's method and let M_u denotes the respective value of the alternative method. Then the gain in dB can be defined as

$$G = 10 \cdot \log_{10} \frac{M_s}{M_u} \tag{12}$$

A positive value of G in the tables suggests that our proposed method yields lower KL divergence than the Sheather–Jones method and is therefore preferable. Both tables show the results for traditional CV that minimises (3). It is clear to see that traditional CV (referred to as CV in the figure legends) is in most cases inferior to the Sheather's method, as it tends to under-smooth the estimates of the distributions. The proposed method for the cost minimisation with the added l_1 term according to (8), was introduced in the last section. The results instantly improve, and our method (referred to as CVL1 in the figure legends) performs on average better than the Sheather–Jones method. The added l_1 term balances out the tendency of traditional CV to under-smooth. In the tables a comparison with a more recent plug-in approach, the Botev's method, is also shown. These results suggest that the proposed CV method including the l_1 term performs better than the Sheather–Jones method or Botev's method. However, it does not hold for all values of N. Our experiments suggest that our proposed

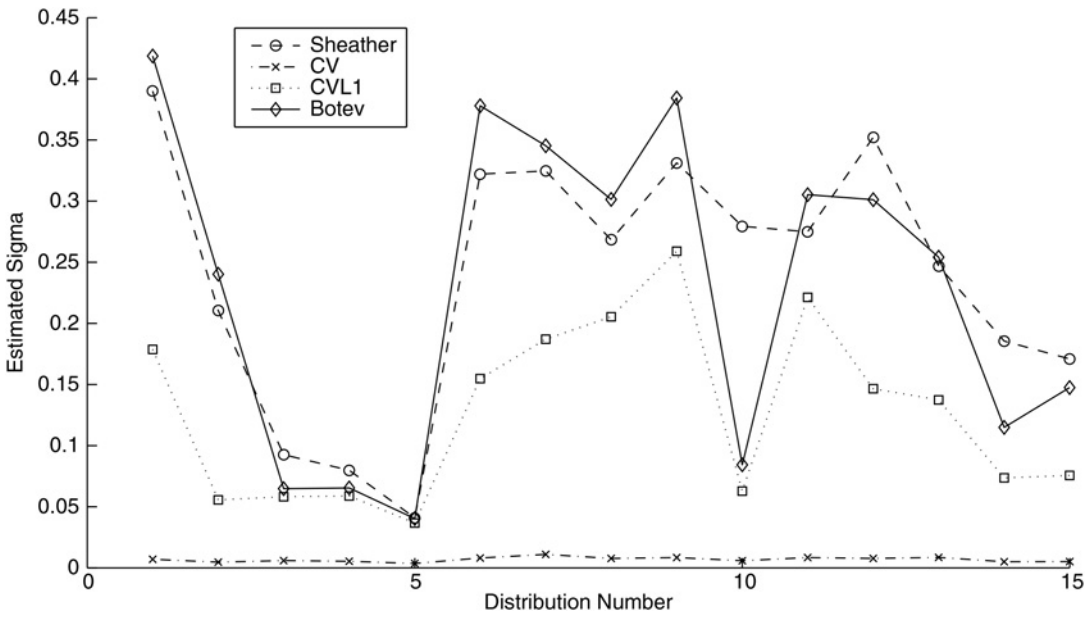

Fig. 2 *Estimated σ for Sheather–Jones, traditional CV, Botev and the proposed method*

IET Signal Process., 2016, Vol. 10, Iss. 3, pp. 280–283

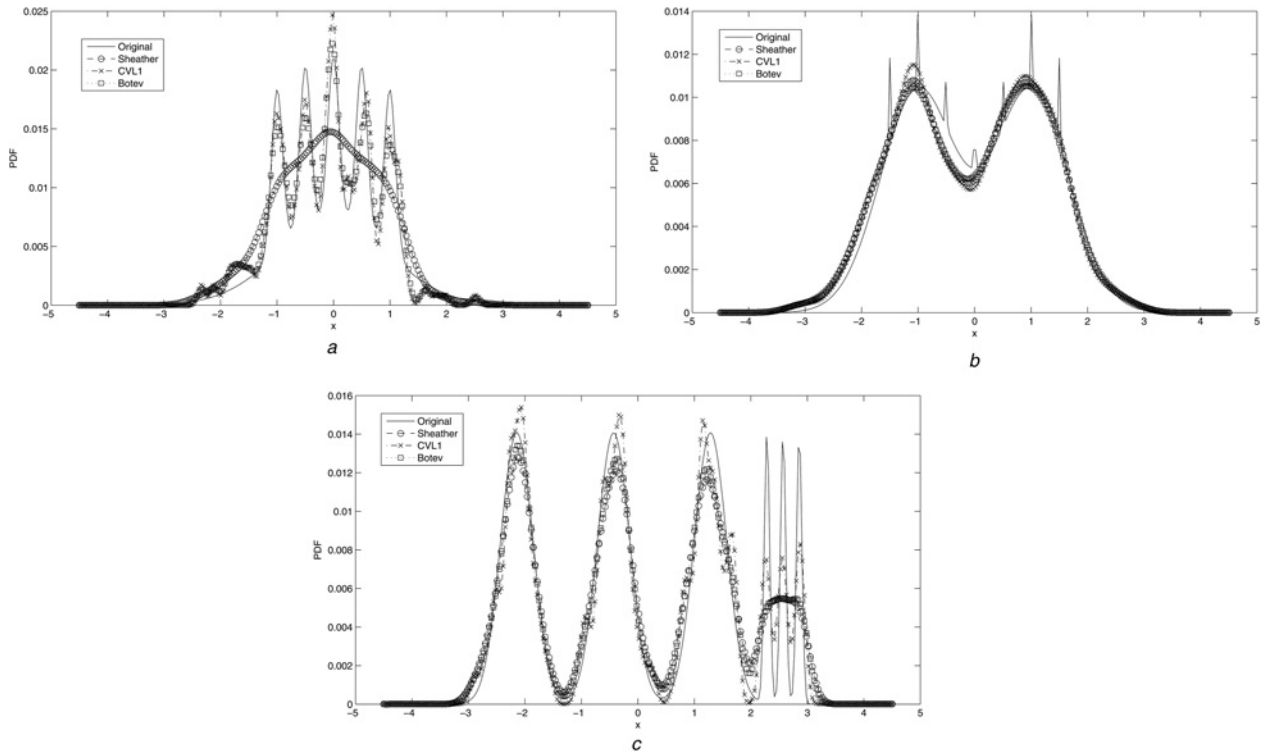

Fig. 3 *Results of KDE for 3 out of 15 example distributions used in this study. Shown are the original, KDE with the Sheather–Jones method, the proposed CV with l_1 norm term method and Botev's method*

a Distribution 10
b Distribution 11
c Distribution 15

methods show better performance than the Botev's method for $N \leq 512$. In this range, the achieved gain is as high as approximately 1 dB. Therefore, the proposed method might be of special practical use if the number of datapoints is limited by the specific application.

Some examples are given for $N = 256$ and one set of data drawn from the 15 example distributions. Our method using the l_1 term with cost minimisation was used for these figures. Fig. 2 shows the estimated σ values for all distributions. In this figure, traditional CV's tendency of under-smoothing is shown, since the estimated σ is much lower than the corresponding bandwidths of the Sheather–Jones method or our proposed method. Our method's bandwidths are consistently smaller than the respective values of the Sheather–Jones or Botev methods, resulting in more detail in the resultant estimates. Fig. 3 shows the density estimates of the proposed methods for three example distributions. Densities 10 and 15 shown in Figs. 3*a* and *c*, respectively, are multi-modal. It is easy to see that for these distributions, the proposed method performs better than the Sheather–Jones method or the Botev method, since it is able to pick up the high frequency content of the distribution. However, for a distribution that includes sharp spikes like distribution 11 shown in Fig. 3*b*, our proposed method performs comparable to the Sheather–Jones method or the Botev method.

5 Conclusion

This paper has shown an alternative way to compute the bandwidth of the kernel for KDE using CV with additional l_1 constraints, thereby largely reducing the under-smoothing effect that traditional CV methods usually exhibit. The proposed method has been compared with the commonly used method by Sheather–Jones and the more recent method by Botev and result in higher fidelity when measured under the KL divergence for a low number of datapoints. The proposed method is especially effective when the distribution to estimate is multimodal and, since it utilises the FFT, is of low algorithmic complexity.

6 References

1 Duda, R.O., Hart, P.E., Stork, D.G.: 'Pattern classification' (Wiley, New York, 2001, 2nd edn.)
2 Silverman, B.W.: 'Density estimation for statistics and data analysis' (Chapman and Hall, 1986), vol. 37, no. 1
3 Marron, J.S., Wand, M.P.: 'Exact mean integrated squared error', *Ann. Stat.*, 1992, **20**, (2), pp. 712–736
4 Jones, M.C., Marron, J.S., Sheather, S.J.: 'A brief survey of bandwidth selection for density estimation', *J. Am. Stat. Assoc.*, 1996, **91**, (433), pp. 401–407
5 Sheather, S.J., Jones, M.C.: 'A reliable data-based bandwidth selection method for kernel density estimation', *J. R. Stat. Soc. B, Methodol.*, 1991, **53**, (3), pp. 683–690
6 Raykar, V.C., Duraiswami, R., Zhao, L.H.: 'Fast computation of kernel estimators', *J. Comput. Graph. Stat.*, 2010, **19**, (1), pp. 205–220
7 Botev, Z.I., Grotowski, J.F., Kroese, D.P.: 'Kernel density estimation via diffusion', *Ann. Stat.*, 2010, **38**, (5), pp. 2916–2957, 10. Available at: http://www.dx.doi.org/10.1214/10-AOS799
8 Loader, C.R.: 'Bandwidth selection: classical or plug-in?', *Ann. Stat.*, 1999, **27**, (27), pp. 415–438
9 Wand, M.P.: 'Fast computation of multivariate kernel estimators', *J. Comput. Graph. Stat.*, 1994, **3**, (4), pp. 433–445
10 Silverman, B.W.: 'Algorithm as 176: kernel density estimation using the fast Fourier transform', *J. R. Stat. Soc. C, Appl. Stat.*, 1982, **31**, pp. 93–99
11 Weisstein, E.W.: 'Fourier transform–Gaussian. From MathWorld – A Wolfram Web Resource', last visited on 22/8/2015. Available at: http://www.mathworld.wolfram.com/FourierTransformGaussian.html
12 Baraniuk, R.G.: 'Compressive sensing', *Lect. Notes IEEE Signal Process. Mag.*, 2007, **24**, (4), pp. 118–120